Hardik Raina

hardikraina079@gmail.com | hardikraina.com | github.com/Raina-Hardik | linkedin.com/in/hardik-raina/

Summary

Senior Machine Learning Engineer with 3+ years shipping production LLM and multimodal AI systems, from fine-tuning and self-hosted model deployment to constrained-generation pipelines and agentic tooling. Published researcher (3 papers), with AI^2 showcased at DAC 2026, and hands-on depth across the full stack: model training through cloud infrastructure (AWS, in-depth).

Education

Jaypee Institute of Information Technology(JIIT) *Noida, UP, India*

B.Tech in Computer Science and Engineering, Specialization in Artificial Intelligence *Sept 2020 - May 2024*

Research Publications

AI^2: EDA Verification Intelligence Platform
Hardik Raina
Industry Showcase, Design Automation Conference (DAC) 2026, 2026

From Data to Decisions: Predictive Analysis Using Novel Approach and Developments on FAIR Prophet
Hardik Raina
Proceedings of the 2024 Sixteenth International Conference on Contemporary Computing (IC3), 2024, Noida, Uttar Pradesh

Language Agnostic Neural Machine Translation: Novel Improvements on Transformer Architecture
Hardik Raina
Proceedings of the 2024 Sixteenth International Conference on Contemporary Computing (IC3), 2024, Noida, Uttar Pradesh

Yoga Pose Detection using VGG
Hardik Raina & Dr. Laxmi Chaudhary
International Conference on Artificial Intelligence, Computing, IoT and Data Analytics (AICTA 2023)', 2023, Chandigarh, India

Work History

Agnisys, Inc. *Multimodal AI, LLMs, RAG, Genetic Algorithms*

Senior Machine Learning Engineer *June 2024 - Present*

- Own **AI^2**, an EDA verification intelligence platform end-to-end: multimodal document understanding over PDF/Docx specs, FSM extraction, SystemRDL/XRSL-driven register-map reasoning, and a RAG pipeline for semantic search over design documentation.
- Trained and fine-tuned LLMs from scratch and explored novel LLM architectures, including a self-hosted, company-wide deployment of a fine-tuned Qwen3.6 model (evaluated against vLLM and Ollama) that replaced OpenAI API dependency for embeddings and synthetic data generation.
- Applied constrained/grammar-based generation to force LLM output into strict UVM-ready JSON schemas for automated verification reporting.
- Used genetic-algorithm-driven fine-tuning strategies for LLMs alongside Reinforcement Learning approaches to IP verification, cutting critical verification turnaround by over 70% and reducing HLD-to-driver generation time from 30 days to under 6 hours.
- Built consumer-facing AI tooling: MCP servers exposing the internal EDA toolchain (IDS-Verify / IDS-Batch) to LLM agents, and a domain-specific agentic coding assistant for EDA workflows with terminal-native VCD waveform rendering.
- Manage AWS infrastructure in depth: autoscaling, S3-based backup and disaster recovery, and cost optimization.

Generative Modeling and Computer Vision *IIM (Indian Institute of Management)*

Research Associate *April 2023 - August 2023*

- Led research and dataset curation, addressing data imbalance in generative models for computer vision.
- Enhanced Recommender System performance by 25.8%, solving cold-start issues using NLP-driven personality matrices.
- Developed a DCGAN to generate high-quality artificial datasets, improving outcomes in vision tasks.
- Increased model refining efficiency by 68% using a Multi-Armed Bandit approach.
- Contributed to a top-performing Recommender System showcased at RecSys 2023.
- Authored reports and publications on generative modeling and AI applications.

Projects

Self-Hosted LLM Inference at Scale

vLLM, AWS EC2, Python/boto3, Docker, Caddy

Production Infrastructure — *2026*

- Serve a 35B-parameter model at full bf16 precision across 4× NVIDIA L40S GPUs under vLLM --- tensor-parallel degree 4, 262K context --- fronted by TLS over both a private tailnet and a public endpoint.
- Cut inference spend by over 50% on a $10.49/hr GPU instance with an idle-stop daemon keyed off vLLM's *monotonic* generation-token counter, so health checks and metrics polling can never be mistaken for real traffic the way instantaneous request gauges are.
- Built single-command cross-region failover in pure boto3: snapshots the model-weights volume, copies it cross-region, relaunches, and reassociates the Elastic IP. Exercised under a genuine outage when the GPU instance type ran out of capacity across all four availability zones in-region.
- Scoped the runner IAM policy to a resource tag rather than instance ARNs so the credential survives failover minting new instance IDs, with explicit denies on terminate, delete, and all IAM actions.

xml-bifurcate

Rust, Tree-sitter, Parallel Compilation

Systems Tool — *2025*

- Built a massively parallel Rust tool that ingests XML intermediate representations of SystemRDL specifications and runs them against a compiler across a thread pool.
- Isolates the exact failing block in a chip design from a single validation run, cutting down manual bisection time.
- Adopted internally at Agnisys as a standard debugging tool for register-map validation.

Image Super-Resolution Service (EDSR)

Python, PyTorch, FastAPI, Docker

Personal Project — *2024 - 2025*

- Implemented and trained a single-image super-resolution model on the EDSR architecture, working from the paper rather than an off-the-shelf checkpoint.
- Measured quality against ground truth with PSNR/SSIM rather than visual inspection, using the metrics to drive architecture and training decisions.
- Productionized the result: Flask to FastAPI, Docker multi-stage builds, health-check endpoints, auto-generated API docs, and an optional Caddy reverse proxy with automatic HTTPS.

Open Source

smelt

Go, ffmpeg

Author & Maintainer — *2026*

- Highly parallel, ffmpeg-powered media transcoding CLI and TUI with a semaphore-gated worker pool and configurable concurrency.
- Hardware-accelerated encoding with automatic GPU probing (NVENC / QSV / VAAPI / AMF / VideoToolbox), falling back to software when unavailable.
- Shipped with CI, versioned releases, and an interactive TUI: live per-file progress, pause/resume, and clean cancellation.

mcat

Skardyy/mcat, 1.3k+ GitHub stars

Open Source Contributor — *2026*

- Diagnosed and fixed three interactive-mode bugs in a 1.3k+ star Rust terminal media viewer: broken Mermaid diagram rendering, a zoom-triggered panic, and broken navigation.
- Navigated an unfamiliar Rust codebase to isolate root causes across rendering and input-handling paths; fixes merged directly by the maintainer as pull request #83.

Skills

Langs	Python, Go, Rust, C/C++, JavaScript/TypeScript, SQL, Lua
ML/AI	LLM fine-tuning & training, Multimodal AI, RAG, Constrained Generation, Genetic Algorithms, Reinforcement Learning, RLHF, PyTorch, Haystack
Domains	Artificial Intelligence, Deep Learning, Generative AI, Self-Hosted Agentic LLM Workflows, MCP Servers, EDA Tooling
Cloud	AWS (IAM, S3, SQS, SNS, autoscaling), Oracle Cloud, GCP, Docker, GitHub Actions
Extra	kuro --- a full embedded-to-mobile IoT stack (Rust/ESP32 firmware, Oracle Cloud proxy, MQTT, Tauri app) shipped solo end-to-end for fun